TPC-DS生成数据

1.首先下载官网 http://www.tpc.org/tpcds/

或者github https://github.com/davies/tpcds-kit.git

2.环境部署：

在Ubuntu12.04上部署，采用apt-get的方式下载所需要的package，因为Ubuntu中yum命令可以会出现源无法使用的问题。如果是centos的操作系统，一般采用yum。

可能出现的问题：

Problem：apt-get时出现”E: Problem with MergeList /var/lib/apt/lists/”错误

Solution：sudo rm /var/lib/apt/lists/\* -vfR 删除apt-get install 的所有软件状态包

sudo apt-get update

Origion: <http://blog.csdn.net/gopain/article/details/13291211>

3. apt-get install gcc g++ flex cmake bison libncurses5-dev

（这是根据一篇网上攻略所写，但不一定都需要）

4.编译

cd DSTtools/tools

cp Makefile.suite Makefile

make

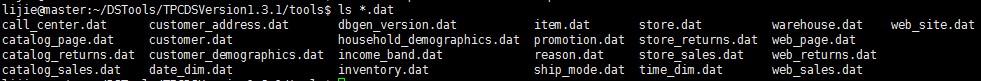
编译成功后，会生成许多.o文件，并且可以执行dsdgen命令

5.生成

dsdgen scale 1 –force

scale指数据规模，1代表1g，-force代表覆盖已有文件

生成25个.dat文件



6.后续

1）需要将tools部署到所有的worker节点上，然后分布式的生成（此处可以参考spark-sql-perf的实现）

2）将生成的数据用于测试，在测试框架上编写测试的workload